Comments on NIST's Draft Profile on Generative AI

Malcolm Murray
Research Affiliate
Centre for the Governance of AI
murray.malcolm@gmail.com

Jonas Schuett
Research Fellow
Centre for the Governance of AI
jonas.schuett@governance.ai

Sam Manning
Research Fellow
Centre for the Governance of AI
sam.manning@governance.ai

Alan Chan
Research Scholar
Centre for the Governance of AI
alan.chan@governance.ai

Leonie Koessler
Research Scholar
Centre for the Governance of AI
leonie.koessler@governance.ai

Markus Anderljung
Head of Policy
Centre for the Governance of AI
markus.anderljung@governance.ai

May 2024

We welcome the opportunity to comment on NIST's Draft Profile on Generative AI (GAI) that complements the AI Risk Management Framework (RMF). We offer the following submission for your consideration and look forward to future opportunities to provide additional input. In February, we also submitted a Response to the RFI Related to NIST's Assignments Under the Executive Order Concerning AI.

About GovAI

The Centre for the Governance of AI (GovAI) is a nonprofit based in Oxford, UK. It was founded in 2018 at the University of Oxford, before becoming an independent research organization in 2021. GovAI's mission is to build a global research community, dedicated to helping humanity navigate the transition to a world with advanced AI. GovAI is part of the Artificial Intelligence Safety Institute Consortium (AISIC).

About the authors

- **Malcolm Murray** is Research Affiliate at GovAI. His research focuses on AI risk management. Before joining GovAI, he was Chief of Research, Risk and Audit, at Gartner, a technology research company. He holds an MBA from INSEAD and is a Superforecaster with Good Judgment.

- **Jonas Schuett** is a Research Fellow at GovAI. His research focuses on the regulation and governance of frontier AI systems, with a special focus on risk management. Before joining GovAI, he advised the UK government on AI regulation and was part of Google DeepMind's Public Policy Team. He has a background in law.

- **Sam Manning** is Research Fellow at GovAI. His research focuses on measuring the economic impacts of frontier AI systems and designing policy options to help ensure that advanced AI can foster broadly shared economic prosperity. Before joining GovAI, he conducted research at OpenAI

and worked on an impact evaluation of a guaranteed income program in the US. He has a MSc in International and Development Economics.

- **Alan Chan** is Research Scholar at GovAI. His research focuses on governing AI agents. He has a background in mathematics and machine learning, and is also a final year PhD student at Mila (Quebec AI Institute).

- **Leonie Koessler** is a Research Scholar at GovAI. Her research focuses on frontier AI regulation and risk management, in particular risk assessment and technical standards. Before joining GovAI, she worked for the German government. She holds a Master of Laws (LL.M.) from King's College London.

- **Markus Anderljung** is Head of Policy at GovAI, an Adjunct Fellow at the Center for a New American Security (CNAS), and a member of the OECD Expert Group on AI Futures. His research focuses on the regulation and governance of frontier AI systems. He was previously seconded to the UK Cabinet Office as a Senior Policy Specialist.

***Note**: The views expressed in this submission are those of the authors and do not represent the views of GovAI.*

Summary

We welcome NIST's effort to create a profile that focuses on the unique challenges of generative artificial intelligence (GAI).[1] The first draft is a commendable first step. In this submission, we suggest further improvements.

Risk list

- The profile should categorize the identified risks in a way that makes the list as actionable as possible, especially in terms of mitigating the risks. We suggest three ways to group the risks: based on the AI lifecycle, the AI triad, or the actors that could contribute to a risk event occurring.
- The risk list could be improved by defining the risks more granularly, clearly specifying the scope of the list, and separating between risks, risk drivers and the absence of safeguards. We provide additional guidance on each of these recommendations.
- We suggest adding the following additional risks: labor automation, persuasion and deception, accidents, collective failure, loss of control, and election interference. For each risk, we also recommend key resources.

Actions

- The profile should categorize the list of actions into core and non-core. Although we appreciate NIST's effort to add a foundational classifier, there are still hundreds of foundational actions. We would recommend adding a core or non-core classifier at the level of the actions themselves.
- The profile should categorize the actions by the level of maturity required in the organization. Some of the recommended actions, including having internal audit functions or establishing risk tolerance levels, are not yet in place at all AI developers. We would recommend organizing the actions in each subcategory by levels of maturity of the organization.
- The profile should categorize the actions by the level of maturity required for the technology. Some actions reference technology that is not yet mature and cannot be implemented immediately, including various identifiers of AI-generated content. We would recommend organizing the actions in each subcategory by levels of maturity of required technology. For key actions, we recommend additional resources.

[1] Note that the NIST Draft Profile, and therefore also this comment, focuses on a particular subset of GAI, namely "generative dual-use foundation models, defined by the Executive Order 14110 as "an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts."

Risk list

Categorizing GAI risks

Risks from GAI systems stem from and can be mitigated by the actions of various actors in the GAI lifecycle, ranging from developers, deployers, users, as well as the rest of society. Here, we focus on risks that could be addressed by AI developers and deployers, i.e. from producing and making available GAI systems. We appreciate NIST's request for feedback on sorting or categorizing the risk list and we agree that this would be useful. Specifically, we believe that the risks should be categorized in a way that makes the list as actionable as possible. Arguably, the most important actions in the risk management process relate to risk mitigation, i.e. actions taken to keep risk at an acceptable level. With that goal in mind, we offer some examples of potential ways to categorize the risks: based on the AI lifecycle, the AI triad, or the actors that could contribute to a risk event occurring. We propose several categorizations rather than one, because different categorizations will be most illuminating for different actors and purposes.

The AI lifecycle

One option is to categorize the risks based on the AI lifecycle (Janjeva et al., 2023). A benefit of this categorization is that the mitigating actions are clearly separated chronologically, taking place during three phases: design, training, and testing of the model; deployment and usage of the model; and longer-term deployment and diffusion of the model. That could result in the following grouping of risks:

Design, training, and testing	Deployment and usage	Longer-term deployment and diffusion
• Environmental (5) • Intellectual Property (9) • Data Privacy (4) • Value Chain and Component Integration (12)	• CBRN Information (1) • Information Security (8) • Confabulation (2) • Dangerous or Violent Recommendations (3) • Obscene, Degrading, and/or Abusive Content (10)	• Human-AI Configuration (6) • Information Integrity (7) • Toxicity, Bias, and Homogenization (11)

Table 1: Categorizing GAI risks based on the AI lifecycle

The AI triad

A second option is to base the categories on the three main inputs to the training of AI models, known as the "AI triad" (Buchanan, 2020). These are data, algorithms, and compute. A benefit of this categorization is that the mitigating actions are clearly delineated by the distinct type of input to the model. That could result in the following groupings:

Data	Algorithm	Compute
• Data Privacy (4) • Intellectual Property (9) • Value Chain and Component Integration (12)	• Confabulation (2) • Human-AI Configuration (6) • Dangerous or Violent Recommendations (3)	• Environmental (5)

• CBRN Information (1) • Information Security (8)	• Toxicity, Bias, and Homogenization (11) • Obscene, Degrading, and/or Abusive Content (10) • Information Integrity (7)

Table 2: Categorizing GAI risks based on the AI triad

Different actors

Finally, a third option is to base the categories on the different types of actors that could contribute to a risk event occurring. A benefit of this categorization is that the mitigating actions can be clearly targeted at the actors contributing to the risk. That could result in the following groupings:

Malicious actors using the model	Non-malicious actors using the model	AI ecosystem actors
• CBRN Information (1) • Intellectual Property (9) • Information Integrity (7) • Information Security (8) • Obscene, Degrading, and/or Abusive Content (10)	• Human-AI Configuration (6) • Toxicity, Bias, and Homogenization (11) • Confabulation (2) • Dangerous or Violent Recommendations (3) • Data Privacy (4)	• Environmental (5) • Value Chain and Component Integration (12)

Table 3: Categorizing GAI risks based on different actors

Guidelines to use when categorizing risks

As noted in the profile, it would be beneficial to organize the 12 risks into categories. There are a few best practices to use when creating a categorized risk taxonomy. Below, we provide some of these as guidelines for NIST to consider when categorizing the risks.

- **Define the risks more granularly.** A risk that is too multifaceted and contains aspects that need to be mitigated in wholly different ways can be hard to manage as a single risk. An example is 6 (Human-AI Configuration) which contains aspects that are quite different in terms of their harm and their mitigation, e.g. automation aversion and deceptive model capabilities.

- **Distinguish risks from an absence of safeguards.** Clearly distinguish aspects that are risks from aspects that are not risks by themselves, but safety measures that are absent or lacking. Currently, some of the risk categories such as 8 (Information Security) contain a combination of aspects of GAI models that create a specific kind of risk (such as helping with offensive cyber capabilities) with aspects that are safety measures that should be adopted (such as ensuring model weights are not stolen).

- **Specify the scope of the list.** The scope of the risk list should be clearly defined. Aspects of scope to specify can include: (1) Type of harm – This can include e.g. casualties, economic damage,

environmental harm, etc. See e.g. the seven types of harm in the UK's National Risk Register (UK Government, 2023). (2) Level of harm – This could be risks with all levels of harm or only risks with significant levels of harm, for example. (3) Causality – One can consider risks from all effects from the development and deployment of a model, or only risks with first-order effects. (3) Time frame – Risks can be limited to those that can have an impact in the next 12 or 24 months or one can consider risks with potential impact further in the future. (4) Region – Risks can be limited to those that are relevant for certain regions, or a global perspective can be taken. (5) AI ecosystem player – One can consider risks for developers and deployers of AI systems, or risks for all players in the AI ecosystem, including users and society at large.

- **Consider whether the risk list should be more mutually exclusive.** Currently, some of the categories show some level of overlap and could be delineated further. If a risk list is used with the purpose of assigning risk owners, i.e. individuals who are ultimately accountable for ensuring the risk is managed appropriately, it can be beneficial to limit overlap between responsibilities. For example, risks 11 (Toxicity, Bias, and Homogenization) and 3 (Dangerous or Violent Recommendations) have some overlap in terms of undesired content output from an AI system. Risk 11 (Toxicity, Bias, and Homogenization) could likely target societal inequality and discrimination risks only.

- **Consider whether the risk list should be collectively exhaustive.** There are potential risks that could be added (see section "GAI risks to consider including" below).

- **Separate risks from "risk drivers".** Risks are typically defined as sources of hazard that can directly lead to harm, without involving other risks. In the current list, some risk categories such as 12 (Value Chain and Component Integration) are not stand-alone risks if that definition is used. For example, if there is "data that has been improperly obtained or not cleaned", then that would exacerbate one of the other risks, rather than creating harm directly. Instead, factors such as value chain and component integration are often classified as "risk drivers" and could be provided in a separate list. Other examples of risk drivers commonly used are race dynamics between companies and limitations in the science of understanding models.

- **Offer multiple risk categorizations.** Different risk categorizations will be most illuminating in different settings. For example, while a compute provider might benefit from the input-based categorization, a developer might benefit from the life-cycle categorization. As such, it seems useful for NIST to offer a range of different risk categorizations.

GAI risks to consider including

We recognize that the risk list is the outcome of a consultation process with many stakeholders in the working group. Although the draft already covers most of the relevant risk areas, we would suggest considering adding the following risks:

Risk	Description	Relevant literature
Labor automation risks	The introduction of new GAI tools can automate tasks and in some cases lead to job loss. Job loss is correlated with a number of negative physical, mental, and social	• Eloundou et al., 2024 • Korinek & Juelfs, 2022 • Korinek & Suh, 2024

	outcomes (Brand, 2015).	• Acemoglu & Restrepo, 2019
Persuasion and deception risks	Risks that GAI systems may manipulate and unduly influence users by exploiting cognitive biases, misrepresenting information, engaging in prolonged and convincing interactions, or deceive humans. This could lead to harmful outcomes like false beliefs, poor decision making, and loss of user autonomy, as well as fraud, election tampering, or losing control of AI systems.	• El-Sayed et al., 2024 • Gabriel et al., 2024 • Anthropic, 2024 • Hubinger et al, 2024 • Scheurer et al, 2023 • Park et al, 2024 • Apollo Research, 2023
Accident risks	Risks driven by unintended model behavior leading to accidents and damages. These risks are distinct from misuse, and often emerge due to poor AI system design and lack of reliability or premature integration into complementary systems.	• Amodei et al., 2016 • Arnold & Toner, 2021 • Maham & Küspert, 2023
Collective failure risks	Risks that may come about from the collective failure of GAI across a wide range of cases, rather than failure from an individual use. For example, suppose a foundation model is biased against a particular group. If that foundation model is widely deployed in hiring, that particular group may be disadvantaged in all of those hiring contexts.	• Bommasani et al., 2022 • Fish et al., 2024 • Dorner, 2021 • Vipra & Korinek, 2023
Loss of control risks	As AI systems become increasingly advanced and autonomous, there is a risk that humans may lose the ability to maintain meaningful control over their development and actions. This could potentially lead to AI pursuing goals misaligned with human values or intentions, making decisions that harm individuals or society, and causing significant economic damage.	• Cohen et al., 2024 • Hendrycks, Mazeika, & Woodside, 2023 • Bengio et al., 2024 • Bernardi et al, 2024
Election interference risk	Risks from AI systems being used to create “deepfakes” or other synthetic video, image, audio, or text content specifically with the goal of influencing political elections.	• Wilder & Vorobeychik, 2019 • Marsden, Meyer, & Brown, 2020 • Buchanan et al., 2021 • Horvitz, 2022 • Park et al., 2023 • Hawes et al, 2023 • Chowdhury, 2024 • AI Digest, 2024

Table 4: GAI risks to consider including

Actions

In terms of actions, we start with general comments and then we proceed to list comments on specific actions.

General comments

- **Categorize the list of actions by the level of maturity of the GAI system developer.** AI risk management is still a nascent field. This means that many aspects of traditional risk management are not yet in place for all developers of GAI systems. For example, many developers do not yet have internal audit functions or established risk tolerance levels. We would therefore suggest organizing the actions in each subcategory by levels of maturity of the developer.

- **Categorize the list of actions by the level of maturity of the technology.** The actions reference many examples of existing technology, such as MG-4.1-004 ("Employ user-friendly channels such as feedback forms, e-mails, or hotlines for users to report issues, concerns, or unexpected GAI outputs to feed into monitoring practices"). These can be implemented straight away. However, there are also references to technology that is not yet as mature and cannot be implemented straight away. For example, MG-4.1-008 ("Integrate digital watermarks, blockchain technology, cryptographic hash functions, metadata embedding, or other content provenance techniques within AI-generated content to track its source and manipulation history") might be dependent on further development. We would therefore suggest organizing the actions in each subcategory by levels of maturity of the referenced technology.

- **Divide the list of actions into "core" and "non-core".** We appreciate NIST's effort to mark some subcategories as "foundational". However, since that foundational tag is at the AI RMF subcategory level, and is applied to the majority of subcategories, that still leaves several hundreds of foundational actions. We would recommend adding a core or non-core classifier at the level of the actions themselves.

- **Make the mapping of risks to actions more consistent**. In the current list, there are some discrepancies in terms of how the actions are mapped to risks. There are examples of actions being mapped to one risk, but they should also be mapped to other risks. For example, MP-2.1-005 ("Review efficacy of content provenance techniques on a regular basis and update protocols as necessary") is mapped to Information Integrity, but should also be mapped to Intellectual Property. There are examples of actions being mapped to risks that are not in the risk categories. MS-1.1-003 ("Conduct adversarial role-playing exercises, AI red-teaming, or chaos testing to identify anomalous or unforeseen failure modes") is mapped to "Unknowns". In general, it is likely worth recognizing that this will in many cases be a many-to-many relationship.

Comments on specific actions

In the following, we outline key aspects that we would argue should feature, or feature more prominently, in the actions. Many of these are commonly accepted and effective practices in other industries; given the potential risks from AI, we think it would make sense to include them. For each action, we provide a description, comments regarding its current treatment, and references to relevant literature.

Govern

Action	Description	Comments	Literature

Capability thresholds	The actions should include the need to specify the level of model capabilities that would cause the developer to pause the development and deployment process.	This is currently mentioned in GV-4.3-003 ("Establish minimum thresholds for performance and review as part of deployment approval ('go/no-go') policies, procedures, and processes, with reviewed processes and approval thresholds reflecting measurement of GAI capabilities and risks"). Given its importance, we would argue this should be given greater prominence.	• DSIT, 2024 • Koessler et al., Risk thresholds for frontier AI, *forthcoming*
Risk thresholds	The actions should include the need to specify the levels of risk that are acceptable, non-acceptable, and acceptable under certain circumstances and with certain actions.	This is currently mentioned in GV-4.3-003 ("Establish minimum thresholds for performance and review as part of deployment approval ('go/no-go') policies, procedures, and processes, with reviewed processes and approval thresholds reflecting measurement of GAI capabilities and risks"). Given its importance, we would argue this should be given greater prominence.	• DSIT, 2024 • Koessler et al., Risk thresholds for frontier AI, *forthcoming*
Safety policy for catastrophic risks	The actions should include the need to establish an overall safety/risk management policy for managing catastrophic risks from GAI. This policy should contain commitments to conduct model evaluations, to pause the development and deployment process if the safety measures are inadequate for a model's level of capabilities, and to verify adherence to the policy.	There is currently no explicit mention of establishing an overall safety/risk management policy. These policies are now in place at Anthropic, OpenAI, and Google DeepMind. Another 13 companies recently committed to producing such policies.	• Anthropic, 2023 • Anthropic, 2024 • OpenAI, 2023 • Google DeepMind, 2024 • DSIT, 2023 • DSIT, 2024
Board risk committee	The actions should include the need to establish a board-level committee that oversees GAI risks, meeting at least quarterly.	There is currently no mention of establishing new board committees.	• Schuett, 2023 • OpenAI, 2024
Chief risk officer	The actions should include the need to appoint a senior executive, who is responsible for all risk management activities, including societal	There is currently no mention of appointing specific senior executives.	• Schuett, 2023

	risks from GAI. The person must not be responsible for product development.		
Central risk function	The actions should include the need for there to be a central risk function (a "second line" in the terminology of the Three Lines Model). The purpose of this function is to provide expert advice and challenge on the management of risk by the organization's leaders.	GV-1.2-001 ("Connect new GAI policies, procedures, and processes to existing model, data, and IT governance and to legal, compliance, and risk functions") refers to risk functions, but it does not specify the need to create a central risk function if one is not present.	• IIA, 2020 • Schuett, 2023
Internal assurance	The actions should include the need for a function that is organizationally independent from senior management, reports directly to the board of directors, and assesses the effectiveness and adequacy of risk management practices.	GV-1.3-003 ("Increase cadence for internal audits to address any unanticipated changes in GAI technologies or applications") and GV-4.1-006 ("Incorporate GAI governance policies into existing incident response, whistleblower, vendor or investment due diligence, acquisition, procurement, reporting or internal audit policies") refer to internal audit, but do not specify the need to establish an internal audit function if one is not present.	• Schuett, 2023a • Schuett, 2023b
External assurance	The actions should include the need for external assurance provided by third parties over the risk management processes and policies and risk and control assessments.	GV-4.1-003 ("Establish policies, procedures, and processes detailing risk measurement in context of use with standardized measurement protocols and structured public feedback exercises such as AI red-teaming or independent external audits") refers to external audits, but only in the limited context of "structured public feedback exercises".	Auditing: • Raji & Buolamwini, 2019 • Raji et al., 2020 • Raji et al., 2022 • Mökander et al., 2021 • Mökander et al., 2023 • Birhane et al., 2024 • Anderljung et al., 2023 Red teaming: • Perez et al., 2022 • Ganguli et al., 2022 • Yong, Menghini, & Bach, 2023 • Rando et al., 2022 • Zhan et al., 2023 • Lermen, Rogers-Smith, & Ladish, 2023

			• Gade et al., 2023 • Anthropic, 2023 Researcher access: • Bucknall & Trager, 2023
Whistleblower protection	The actions should include the need for there to be policies protecting whistleblowers to increase transparency.	GV-4.1-006 (“Incorporate GAI governance policies into existing incident response, whistleblower, vendor or investment due diligence, acquisition, procurement, reporting or internal audit policies”) refers to whistleblower policies, but only in the limited sense of incorporating GAI governance policies into existing policies. Given the differences between GAI risks and other risks, dedicated whistleblower policies might be necessary.	–
Risk culture	The actions should include the need for there to be policies and practices that encourage and enable a robust risk culture, including organization-wide awareness of risks, controls and risk tolerances.	GV-1.3-005 (“Reevaluate organizational risk tolerances to account for broad GAI risks, including: Immature safety or risk cultures related to AI and GAI design, development and deployment, public information integrity risks, including impacts on democratic processes, unknown long-term performance characteristics of GAI”) references risk cultures, but only in the limited sense that risk tolerances should be reevaluated to account for immature GAI risk culture. This does not include the actual creation of a robust risk culture.	• Manheim, 2023
Emergency response / Crisis management	The actions should include the need for there to be policies established for emergency response/crisis management. These should include scope, roles and responsibilities, response procedures, communication protocols, resources, training and monitoring.	MP-3.4-008 (“Involve the end-users, practitioners, and operators in AI system prototyping and testing activities. Make sure these tests cover various scenarios where content provenance could play a critical role, such as crisis situations or ethically sensitive contexts”) references crisis situations, but	–

		only in the limited context of content provenance.	

Table 5: Comments on specific actions in the Govern function

Map

Action	Description	Comments	Literature
Risk identification process	The actions should include the need for there to be a systematic process for identifying new risks. This should be regularly reviewed for its adequacy and updated if necessary.	GV-4.2-012 ("Include relevant AI Actors in the GAI system risk identification process") references there being a risk identification process, but the actions don't specify explicitly that the risk identification process will need to be revisited and updated to accommodate GAI risks.	• Koessler et al, 2023
Risk register	The actions should include the need for there to be a risk register that contains a list of all identified risks and the results of previous assessments. This should be regularly updated and used for decision-making regarding risk prioritization and mitigation.	The actions do not currently include any references to establishing a risk register should one not exist or to update it in light of GAI risks.	• IEC 31010:2019
Continuous risk monitoring	The actions should include the need for there to be a process for continuous monitoring of risks. This process should measure changes in the levels of risk.	MS-2.2-006 ("Implement continuous monitoring of GAI system impacts to identify whether GAI outputs are equitable across various sub-populations. Seek active and direct feedback from affected communities to identify issues and improve GAI system fairness") mentions continuous monitoring, but only in the context of one type of risks - fairness and inequality. It is also needed for other risks, that are arguably faster moving.	• Brundage et al, 2022
Emerging risk monitoring	The actions should include the need for there to be a process specifically for monitoring emerging risks.	The actions do not currently include any mentions of tracking emerging risks.	–

Table 6: Comments on specific actions in the Map function

Measure

Action	Description	Comments	Literature
Model evaluations	The actions should include the requirement to do evaluations of all capabilities of models that can lead to harm.	The actions cover model evaluations fairly extensively, e.g. in MS-2.3-005 (“Evaluate claims of model capabilities using empirically validated methods”) and MS-2.5-002 (“Avoid extrapolating GAI system performance or capabilities from narrow, nonsystematic, and anecdotal assessments”).	• Liang et al., 2021 • Srivastava et al., 2022 • Shevlane et al., 2023 • Kinniment et al., 2023 • Phuong et al., 2024 • Weidinger et al., 2024 • Chan, 2024
Red-teaming	The actions should include requirements for extensive red-teaming so that models are tested by teams of different expertise and aims.	The actions cover model evaluations extensively, e.g. in MS-1.1-003 (“Conduct adversarial role-playing exercises, AI red-teaming, or chaos testing to identify anomalous or unforeseen failure modes”) and MS-2.7-016 (“Perform AI red-teaming to assess resilience against: Abuse to facilitate attacks on other systems (e.g., malicious code generation, enhanced phishing content), GAI attacks (e.g., prompt injection), ML attacks (e.g., adversarial examples/prompts, data poisoning, membership inference, model extraction, sponge examples)”).	• Perez et al., 2022 • Ganguli et al., 2022 • Yong, Menghini, & Bach, 2023 • Rando et al., 2022 • Zhan et al., 2023 • Lermen, Rogers-Smith, & Ladish, 2023 • Gade et al., 2023 • Anthropic, 2023
(Semi-) quantitative risk estimates	The actions should include the requirements to estimate the impact and likelihood of key risk events following a quantitative or semi-quantitative approach. The estimates should be used to inform high-stakes development and deployment decisions (e.g. whether to release a model).	The actions make reference to risk estimates in e.g. MP-5.1-008 (“Prioritize risk acceptance, management, or transfer activities based on risk estimates”). However, the actions do not specify if the estimates should be qualitative or quantitative.	• Schuett et al., How to estimate the impact and likelihood of risks from AI, *forthcoming*
Risk modeling	The actions should include the requirements to do extensive modeling of risk scenarios, that is, pathways from risk factors to harm (also referred to as “threat assessment”). Risk modeling should be conducted to	The actions make reference to risk modeling in e.g. MS-2.13-001 (“Leverage domain expertise when modeling complex societal constructs such as toxicity”). However, the actions do not prescribe modeling for all relevant risks.	• Anthropic, 2023 • Anthropic, 2024 • OpenAI, 2023 • Google DeepMind, 2024

	better understand risk factors, help with producing (semi-) quantitative risk estimates, and identify where risks can be best managed.		

Table 7: Comments on specific actions in the Measure function

Manage

Action	Description	Comments	Literature
Staged release	The actions should include requirements to consider different ways to release the model, for example leveraging a staged release approach.	The actions do not currently include recommendations for different approaches in releasing models.	• Solaiman et al., 2019 • Solaiman, 2023 • Shevlane, 2022 • Bucknall & Trager, 2023 • Partnership on AI, 2023
Access and usage restrictions	The actions should include requirements specifying that appropriate access and usage restrictions should be in place to manage risks to their relevant risk thresholds.	The actions make reference to these types of controls in e.g. MS-2.7-007 ("Identify metrics that reflect the effectiveness of security measures, such as data provenance, the number of unauthorized access attempts, penetrations, or provenance verification") and MS-2.6-010 ("Verify that systems properly handle queries that may give rise to inappropriate, malicious, or illegal usage, including facilitating manipulation, extortion, targeted impersonation, cyber-attacks, and weapons creation").	• O'Brien, Ee, & Williams, 2023 • DSIT, 2023 • OpenAI, 2024
Cybersecurity	The actions should include requirements to have cybersecurity controls that are appropriate for the level of risk and capabilities of a given model. This includes controls focused on e.g. Model weights and Algorithmic insights.	The actions make reference to cybersecurity controls in MS-2.2-007 ("Implement robust cybersecurity measures to protect both the research data, the GAI system and its content provenance from unauthorized access, breaches, or tampering and unauthorized disclosure of human subject information"). Given the importance of cybersecurity, they should feature more prominently.	• Nevo et al., 2024 • Anthropic, 2023

Table 8: Comments on specific actions in the Manage function